A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records

Nuno Freire, José Borbinha, Pável Calado, Bruno Martins
IST / INESC-ID
Apartado 13069,
1000-029 Lisboa, Portugal
{nuno.freire, jlb, pavel.calado, bruno.g.martins}@ist.utl.pt

ABSTRACT

This paper describes an approach for performing recognition and resolution of place names mentioned over the descriptive metadata records of typical digital libraries. Our approach exploits evidence provided by the existing structured attributes within the metadata records to support the place name recognition and resolution, in order to achieve better results than by just using lexical evidence from the textual values of these attributes. In metadata records, lexical evidence is very often insufficient for this task, since short sentences and simple expressions are predominant. Our implementation uses a dictionary based technique for recognition of place names (with names provided by Geonames), and machine learning for reasoning on the evidences and choosing a possible resolution candidate. The evaluation of our approach was performed in data sets with a metadata schema rich in Dublin Core elements. Two evaluation methods were used. First, we used cross-validation, which showed that our solution is able to achieve a very high precision of 0,99 at 0,55 recall, or a recall of 0,79 at 0,86 precision. Second, we used a comparative evaluation with an existing commercial service, where our solution performed better on any confidence level (p<0,001).

Categories and Subject Descriptors

E.1 [Data] Data Structures – Records; H.3.7. [Digital Libraries]; I.2.7 [Artificial Intelligence]: Natural Language Processing - Text analysis; I.7.m [Document and Text Processing]:[Miscellaneous]

General Terms

Algorithms. Documentation Experimentation.

Keywords

Information extraction, entity recognition, entity resolution, metadata, geographic information.

1. INTRODUCTION

A wide range of potentially usable business information exists in unstructured forms. Although machine readable, those scenarios might follow data models with generic semantics, or may not follow any data model at all, when it consists of natural language texts (it was even estimated that 80% to 90% of business information may exist in those unstructured forms [1][2]).

As society becomes more data oriented, much interest has arisen in these unstructured sources of information. This interest gave origin to the research field of *information extraction*, which looks for automatic ways to create structured data from unstructured data sources [3]. An information extraction process selectively structures and combines data that is found, explicitly stated or implied, in texts or data sets with semi-structured schemas. The final output of the extraction process will vary according to the purpose, but typically it will consist in semantically richer data, which follows a more structured data model, and on which more effective computation methods can be applied.

A particular scenario of information extraction deals, for example, with the references to geographic entities. In many cases, geographic entities are represented in data sets with place names occurring in natural language expressions, instead of using, for example, structured attributes with geospatial coordinates.

This scenario deals with two particular problems: how to locate, in the source data, these references (place name recognition) and how to resolve the references to an exact geographic entity (place name resolution). The typical output of this scenario is the identification of the references within the source data, and their resolution to geospatial coordinates or to an identifier in a set of disambiguated geographic entities, such as within a geographic gazetteer like Geonames [4]. This scenario is commonly referred to as geoparsing.

Digital libraries frequently have structured data records associated with the digital resources they hold. These data, commonly referred in the digital library community as *metadata*, may serve many purposes, but one of the most relevant is resource discovery. With this purpose in mind, metadata records often contain unstructured information in natural language texts that might be useful for helping the user to evaluate if a particular resource is relevant for his information need. These data also can be indexed, so that users also can search them. In this scenario it is evident that if the geographic entities can be represented in structured models based on geospatial coordinates, the user experience is expected to improve.

Several recent research efforts within the information retrieval community have addressed the general topic of geographic text analysis, for instance by developing versatile and comprehensive methodologies for mapping natural language expressions, given

JCDL'11, June 13–17, 2011, Ottawa, Ontario, Canada.
Copyright 2011 ACM 978-1-4503-0744-4/11/06...$10.00.

over textual documents, describing locations, orientations and paths, into the geographic entities they refer to [5][6][7].

However, in metadata records, we find a scenario where unstructured text exists within structured data. This provides a richer context that can be used for supporting the recognition and resolution of place names within the unstructured text. This was the hypothesis tested in our work.

The remainder of this paper will follow, in Section 2, by describing other work related to the problem we addressed. Section 3 follows with a description of our general approach. Section 4 describes the implementation of that approach in a system we called of Metadata Geoparser. Section 5 presents experimental results. Section 6 concludes.

2. RELATED WORK

Place name recognition concerns the delimiting, in unstructured text, of the character strings that refer to place names. This is a particular instance of the more general problem of Named Entity Recognition (NER), which has been extensively studied in the Natural Language Processing (NLP) community. Place name resolution refers to associating the recognized references into the corresponding entries in a gazetteer (a dictionary of geographic features). This latter sub-task has been addressed by the Geographic Information Retrieval (GIR) community.

2.1 Named entity recognition

The NER task, as proposed by the NLP community, refers to locating and classifying atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of time, quantities, monetary values, percentages, etc. [8][9]. Current state-of-the-art solutions can achieve near-human performance, effectively handling the ambiguous cases (e.g., the same proper name can refer to distinct real world entities), and achieving F-scores around 0,90. A recent survey of this area is available in [9].

Initial approaches, which are nonetheless still commonly used, were based on manually constructed finite state patterns and/or collections of entity names [9]. In general, these pattern-based approaches attempt to match against a sequence of words in much the same way as a general regular expression matcher. However, named entity recognition is considered to be a typical scenario for the application of machine learning algorithms, because of the potential availability of many types of evidence, which form the input variables for the algorithms [14].

A major factor supporting the use of machine learning algorithms for entity recognition is their capacity to adapt to each case. Thus, they can be deployed with greater flexibility on distinct corpora from different domains, languages, etc. Different types of text analysis methods make available several types of evidence on which to base the named entity recognition systems. Still, not all evidences will be present in every corpus, and not all text analysis techniques will be able to identify the same types of evidence, so the capacity of machine learning algorithms to adapt to each case, makes it a very good solution for entity recognition. This analysis is supported by the rising trend in usage of machine leaning in this research area [9].

Two particular types of supervised machine learning algorithms have been successfully used for entity recognition. Early applications applied classification algorithms, which basically classify words, or groups of words, according to their entity type. Some examples are Support Vector Machines [10], Maximum Entropy Models [12] and Decision Trees [11].

Nevertheless, in entity recognition, as in other natural language related tasks, the problem was shown to be better solved with sequence labeling algorithms. The earliest sequential classification techniques applied to entity recognition were Hidden Markov Models [13]. However this technique does not allow the learning algorithm to incorporate into the predictive model the wide range of evidence that is available for entity recognition. This limitation has lead to the application of other algorithms such as the Maximum Entropy Markov Model [14] and Conditional Random Fields [15]. Conditional Random Fields is currently the technique that provides the best results for entity recognition. It has sequence classification learning capabilities together with the flexibility to use all the types of evidences that entity recognition systems can gather [16].

2.2 Place name resolution

While NER approaches can be designed to rely entirely on features internal to the documents, place name resolution requires always external knowledge for translating place names into geospatial footprints. Geonames [4] is an example of a modern wide-coverage gazetteer, describing over 6,5 million unique places from all around the world and having been used in many GIR experiments [18]. Wikipedia and related services such as DBPedia [19], a service leveraging on structured information extracted from Wikipedia and published as Linked Data on the Web, are also increasingly reliable sources of geographical information. For instance, DBPedia describes more than four hundred thousand geographic locations.

Similarly to the general case of named entity recognition, the main challenges in resolving place references are related to ambiguity in natural language. Ambiguity problems can be characterized according to two types, namely geo/non-geo or geo/geo [20]. Geo/non-geo ambiguity refers to the case of place names having other non geographic meanings (e.g., Georgia may refer to the country or to a person). Some common words are, for instance, also place names (e.g., Turkey). On the other hand, geo/geo ambiguity arises when two distinct places have the same name. For instance, almost every major city in the Europe has a sister city of the same name in the Americas. The geo/non-geo ambiguity is addressed when identifying mentions to places over text, while geo/geo ambiguity is addressed while disambiguating the recognized place references.

Several approaches for place reference resolution have been proposed in the past. For instance, in [18], a system to resolve locations mentioned in transcripts of news broadcasts is described. The authors report matching 269 out of 357 places (75%) in the considered test segments, while using a gazetteer of about 80,000 items. In [21], the author reports a value of 96% precision for geographic name disambiguation in Japanese text, with a gazetteer of 55,000 Japanese and 41,000 foreign names. A variety of approaches for handling place references in textual documents have been surveyed in [5], where it is concluded that most methods rely on gazetteer matching for performing the identification, together with natural language processing heuristics for performing the disambiguation. Some often used heuristics are default senses (i.e., disambiguation should be made to the most important referent, estimated with basis on population counts), and geographic heuristics such as the spatial minimality (i.e., disambiguation should minimize the bounding polygon containing all candidates referents).

Despite the technology being very recent, commercial services offering geoparsing functionality are available. Metacarta, a commercial company focused on GIR technology, provides a freely-available Web service that can be used to recognize and disambiguate place references over text [22]. Another commercial example is the Yahoo! Placemaker Web service[1].

3. GENERAL APROACH

Previous research on place name recognition has focused mainly on natural language processing, involving text tokenization, part-of-speech classification of the words, word sequence analysis, etc. It is by reasoning on the output of this process that references to places are recognized. Recognition is therefore dependent of the evidence given by the natural language text to identify a reference to a place.

In previous work [23][27], we observed that current entity recognition techniques underperform, when applied to digital library metadata. Our analysis pointed us to three limiting factors:

- Values in the data attributes may not contain enough lexical evidence to support the recognition and resolution of entities using natural language processing. In many cases, the data consists in short expressions, which are very poor in the basic units of meaning of natural languages (e.g. words, their part-of-speech class, expressions, etc.);
- The language of the text within data values may be hard to determine, either because it is not clearly stated or because automatic language guessing techniques have poor performance on very small texts;
- Entity recognition systems have been used as black boxes that recognize entities from text. When applied to data records, where the text is extracted and provided as input to the system, the entity recognition process does not take advantage of the data structure and all the evidence that it can provide.

The contextual evidence in the structure of the metadata records may compensate for the lack of lexical evidence, contributing to improve results in a way similar to the successful use of corpus and document level evidence on previous work on entity recognition [24][25][26]. Thus, the place name resolution step should also benefit from the data structure, particularly when resolution is to be done on data sets with rich semantics, such as Geonames [4].

In order to exploit the structure of the data, we designed an approach that integrates support for structured data throughout the complete recognition and resolution process. Figure 1 shows the process of the traditional black box approach. The first task consists in selecting the text within the metadata record, where place names are to be recognized and resolved. The rest of the recognition and resolution process is performed only on the text, ignoring any other information from the metadata record.

The next step consists in basic text processing to transform the source text into a series of paragraphs, sentences and finally, tokens (words, punctuation marks, numbers, etc).

The text tokens are then analyzed and associated with any evidence that may support the recognition and resolution of the place names. Typically, it includes lexical analysis of sentences (e.g. part-of-speech tagging), analysis of words features (e.g. capitalization), and checking the existence of the tokens in collections of entity names (e.g. gazetteers).

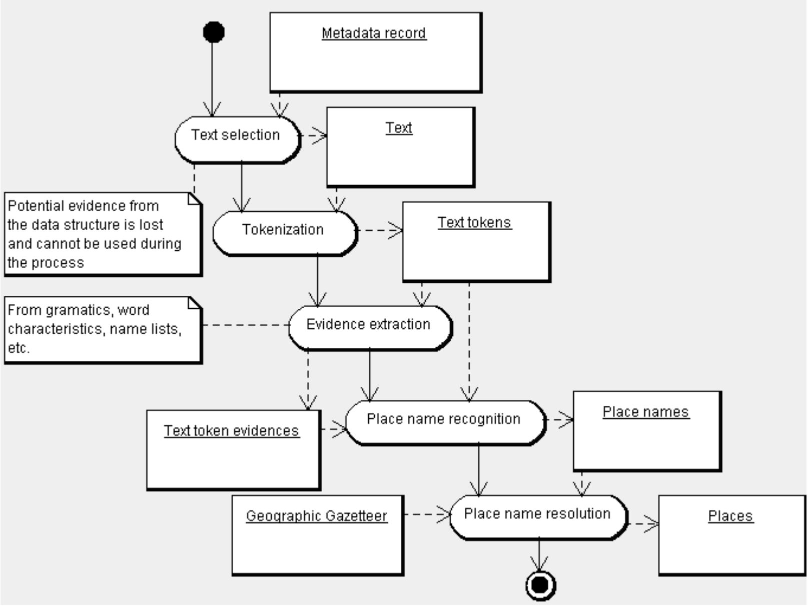

Figure 1 – Black box application of entity recognition and resolution to semi-structured data

[1] http://developer.yahoo.com/geo/placemaker/

Table 1 - Comparison to the proposed approach against the black box approach

Activity	Black box approach	Proposed approach
Text selection	The text is selected from the relevant data fields, and passed forward. The context provided by the data structure is not available for the activities that follow.	The relevant data fields are selected and passed forward. These fields, and the whole record, are available throughout the process.
Tokenization	Performed similarly in both approaches.	
Evidence extraction	Evidence is extracted from the text tokens.	The metadadata field itself provides extra evidence. Also extra evidence can be extracted from structured fields on the same record.
Place name recognition	Can be performed by machine learning, dictionary based, or rule based techniques.	Can be performed with the same techniques, but with more evidence available. Since the text typically consists in short sentences, reasoning techniques used should not be dependent on grammatical evidence.
Place name resolution	Can be performed by machine learning or rule based techniques.	Performed with the same techniques, but with more evidence available.

The following step is the recognition of the place names. It consists in reasoning on the sequence of tokens and the associated evidence. The outcome consists in the recognized names of places and their respective positions in the source text.

The final step is again a reasoning task for resolution of the recognized place names, to specific places described in the target gazetteer.

In our approach, the same general steps are executed. However, they are executed with the metadata record always available, from where further evidence may be used. We also propose that the choice of techniques should be appropriate for the characteristics of the text in metadata records. Table 1 highlights the main changes of executing the process according to our approach.

4. IMPLEMENTATION

The implementation of our approach resulted in a system called the Metadata Geoparser (MG). From the state of the art techniques of named entity recognition and place name resolution, we have chosen a combination of them that would better fit to the characteristics of the metadata. Our system was designed to be language independent, without independence of lexical evidence, and adaptable to different metadata formats. So at its core the MG system uses machine learning algorithms, supported by a wide variety of types of evidence.

This section will first describe the architecture of the MG system and follow with the details of the execution of a process of recognition and resolution of place names.

4.1 Architecture

The main software components of the MG system are depicted in Figure 2. Since in metadata records the language of the text can be uncertain or hard to guess, due to containing very short sentences, the basic text processing and tokenization is done according to the language independent word breaking rules of UNICODE[2], used by the *TextTokenizer* component.

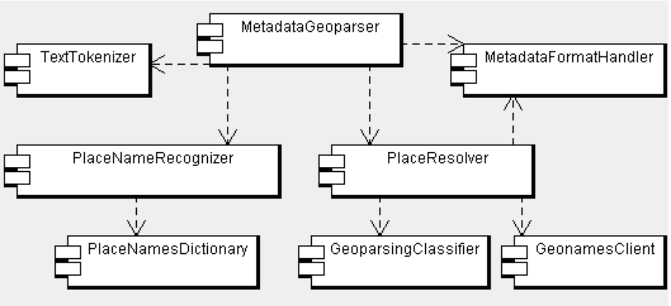

Figure 2 - Metadata Geoparser components diagram

A language independent approach was also chosen for the method of place name recognition. Recognition is done by the *PlaceNameRecognizer* component, based on sequential lookups of the text tokens in a collection of known place names. Therefore recognition is performed without using any lexical evidence (although some lexical evidence is gathered for later use in the process). This collection of place names was created from all the names, and alternative names, of all places described in the Geonames gazetteer. To ensure the performance of the name lookup operations, the place names were indexed in a B-tree [28]. We also use a collection of person names indexed in a separate B-tree. This collection is used to provide evidence, which may help in the decision of not recognizing the name as a place, since it may actually be a reference to a person. In addition, the *PlaceNameRecognizer* component extracts some evidence about the recognition of the name for later reasoning.

The system is designed to be as independent of the metadata format as possible. Any operation dependent of the metadata format in use is abstracted by an interface, implemented by the *MetadataFormatHandler* component. Current implementations include support for Dublin Core [29] and Europeana Semantic Elements[3].

[2] http://www.unicode.org/reports/tr29

[3] http://group.europeana.eu/c/document_library/get_file?uuid=a83 0cb84-9e71-41d6-9ca3-cc36415d16f8&groupId=10602

The *PlaceResolver* component is responsible for extracting any further evidence that can be used for reasoning. Evidence can also be gathered from the metadata record, including other fields besides those where the existence of place names is being analyzed. The *PlaceResolver* uses the *GeonamesClient* component for searching in Geonames. The *GeonamesClient* uses the Geonames web services interface for finding all possible places with a recognized name, or a similar name.

The *MetadataFormatHandler* provides data from the metadata records to the *PlaceResolver*, which will compare it with data obtained from Geonames, in order to support the resolution of the place names.

The ultimate decision to recognize a place name and its resolution are also done by the *PlaceResolver*, through a machine learned classifier component. The decision to use a classifier instead of a sequence label classifier was based on the fact that we wanted the system to be independent of language and lexical evidence, and also due to the predominance of short sentences in metadata records.

Several machine leaning classification algorithms were evaluated and compared, including Support Vector Machines [10], Maximum Entropy Models [12], Decision Trees [11] and Bayesian networks [17]. We have chosen to use a Random Forest [30] classifier for the MG system, since it consistently resulted in higher results in the F_1-measure and lower mean absolute error, for cross-validation tests. The Random Forest algorithm was configured for ten trees of five features. All classification algorithms were tested with the same procedure discussed in Section 5.2.

More details about the role of each component in the execution of a place name recognition and resolution process are provided in the following section.

4.2 Recognition and resolution execution

The execution of a metadata geoparsing request is depicted in Figure 3. In the first step, the *MetadataFormatHandler* selects the fields of the record where place names should be recognized. For example, from the Dublin Core elements, the following are analyzed:

- dcterms:spacial
- dc:coverage
- dc:title
- dcterms:alternative
- dc:subject
- dc:description
- dcterms:tableOfContents

A different choice of the fields could have been made. The choice should be based on the objectives of the information extraction task. The above selection of fields was done with information retrieval in mind.

Each field is tokenized and place names are recognized by name lookup of the word tokens. When a place name is found, some information is associated with it for later usage as evidence for resolving the name. The evidences gathered at this stage are:

- An identifier, such as an URI, of the metadata field;
- If the name was found in the beginning of a sentence;
- The length of the name in characters;
- The number of words in the name;
- The case of the name. If the name was found in lowercase, uppercase, or with only the first letter in uppercase;
- If the name was found in the middle of a sequence of

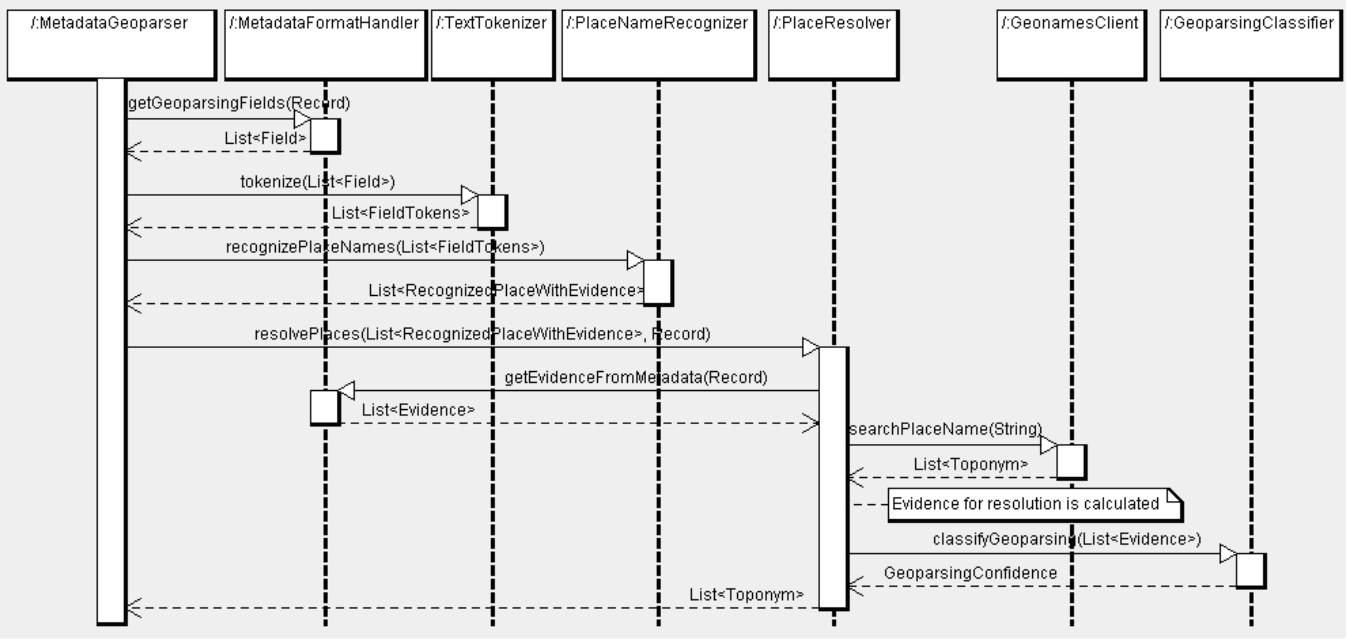

Figure 3 – Typical sequence diagram of a metadata geoparsing request

tokens, which all have the first letter in uppercase;

- The token that immediately preceded the place name;
- If the previous token was found in a collection of person names;
- If the place name was also found in a collection of person names.

The recognized place names, their associated evidence, and the complete metadata records are then processed by the *PlaceResolver* component. The first step at this stage is to find evidence in the metadata record, which may support the recognition and resolution of the place names. Our implementation supports two types for this kind of evidence:

- Record country provenance – the country of origin of the metadata record. The Dublin Core *MetadataFormatHandler* will try to identify the country from a *dcterms:provenance* field.
- Record and resource languages – The Dublin Core *MetadataFormatHandler* will try to identify the languages from *a dcterms:language* field.

The next task performed by the *PlaceResolver* is to find all possible candidates for the resolution of a recognized place name to a place described in Geonames. The place name is used to query Geonames via its web services interface, which then returns places with the same or a similar name.

On the next step, further evidence is gathered for each of the possible resolutions of each place name. The following evidence is gathered:

- Country and country of origin comparison – indicates if the resolution candidate is located in the same country as the origin of the metadata record;
- Number of other places also from the same country found in the same record;
- Place name comparison – Indicates if the name matched the main name of the place, or if it matched a variant name.
- Languages matched – indicates if the matched name was in the same language as the metadata record or resource;
- The shortest edit distance between the name of the entity and the names, and variant names, of the resolution candidate;
- The sum of the geographical distance to the other places found in the same record. The distances are calculated from the coordinates of the places' geographical centre.
- The type of geographic feature (Continent, Country, Administrative area, city, etc.);
- The population of the place.

This final decision to choose a resolution candidate for each recognized place name is made by the *GeoparsingClassifier*, by reasoning on the gathered evidence, and classifying it as "match" or "non-match". The classifier outputs the probability of each of the resolution candidates being the correct one. The one with the highest probability is chosen and returned in the final answer. If none of the resolution candidates achieves a minimum probability threshold for the class "match", the place name is considered not to be referring to a place, so it is ignored and not included in the results.

5. EVALUATION

The evaluation of our approach was performed in the data sets from Europeana[4], which consist in descriptions of digital objects of cultural interest. This dataset follows a data model using mainly Dublin Core[5] attributes, structured in a schema named Europeana Semantic Elements[6]. In this data set, place names appear in data fields for titles, textual descriptions, tables of contents, and subjects.

The dataset contains records originating from several European institutions from the cultural sector, such as libraries, museums and archives. Several European languages are present, even within the description of the same object, for example when the object being described is of a different language than the one used to create its description. Institutions from where this data originates follow different practices for describing the digital objects, which causes the existence of highly heterogeneous data. Lexical evidence is very limited in this data set, so it provides a good scenario for the evaluation of the evidence made available by the structure of the data.

The results of the MG system were evaluated by two methods. First, a cross-validation test aimed to prevent the system from over fitting the training data. And second, to compare the results against those obtained by another existing similar service, namely Yahoo! Placemaker.

5.1 The evaluation data set

An evaluation of the MG system was performed on a selected collection of metadata records from Europeana. This collection was created by selecting records containing place names that exist in the Geonames gazetteer. In order to have heterogeneous data, only up to 20 records were selected from each data provider, yielding a total of 752 records. This allowed the collection to have diverse languages and metadata with different characteristics.

The evaluation collection was processed by two Geoparsers, the MG system and Yahoo! Placemaker service (described in Section 5.3 - Comparative evaluation of the Metadata Geoparser). The results of both were inspected and annotated by a person. The manual annotation was sometimes uncertain, because the metadata record may not contain enough information to enable a person to do the annotations. For example, some sentences with place names were too small and no other information was available in the metadata record to support a decision. Annotation was performed as follows:

- Place names were annotated as *location* and include the identifier (an URI) of the place in Geonames;
- If the annotator was not sure if a mentioned entity was a place, he would annotate it as unknown. These annotations were not considered for the evaluation of the results;

[4] http://www.europeana.eu/

[5] http://dublincore.org/

[6] http://group.europeana.eu/c/document_library/get_file?uuid=a830cb84-9e71-41d6-9ca3-cc36415d16f8&groupId=10602

- If the annotator was sure that a mentioned entity was a place, but could not know for sure how to disambiguate it, it would be annotated as *location_unknown*. These annotations were used only for testing the entity recognition phase, and were not considered for the evaluation of the final results;
- If, for any place name, the annotator was able to identify the country of the place but was not able to disambiguate between alternatives within that country, he would use his own judgment to choose the most likely place (the most populated, for example). Although these annotations seam uncertain, they mimic the desired behavior for the Metadata Geoparser, so they were considered for the evaluation of the results;
- If a place name was not being used to refer to a place (for example, if it is an ambiguous word, or is the name of a person) it would be annotated as *not_location*. These annotations were considered for the evaluation.

In total, the 752 records of the evaluation collection contain 2823 annotated place names.

5.2 Evaluation of the machine learned classifier

The reasoning component of the MG system was trained on the evaluation collection. In order to prevent it from over fitting to the training data, we performed a cross-validation test.

The cross-validation test involves partitioning the evaluation collection into complementary subsets, testing the classifier on one subset, while training on the remaining subset. Ten-fold cross-validation was performed using different partitions, and the validation results were averaged over the ten runs.

For this evaluation, the following measurements were taken:

- Precision: the percentage of correctly identified places in all places found.
- Recall: the percentage of places found compared to all existing places in the data.
- F_1-measure: the weighted harmonic mean of precision and recall.

The MG system only recognizes place names if the probability given by the Random Forest classifier for the class "match" is higher than a minimum threshold (the minimum confidence level). The measurements were taken at five levels of minimum confidence, and they are shown in Figure 4.

We consider the results to be a positive indication that the predictive model is not over fitting the training data. The measured values for the Mean Absolute Error (0,16) and Root Mean Squared Error (0,26) also support this analysis.

During the manual annotation process we noticed the high level of uncertainty of geoparsing. Therefore, we consider the results for precision and recall to be very positive.

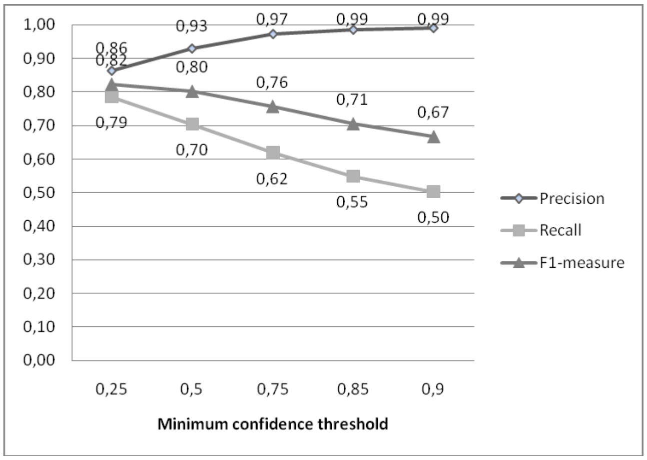

Figure 4 – Measured precision, recall and F_1-measure, using 10-fold cross-validation

The system is able to achieve very high precision of 0,99 at 0,55 recall, or reach a recall of 0,79 at 0,86 precision. These results show that, when using the MG system, we need to carefully choose a minimum confidence level, since to obtain near 1,0 precision, recall lowers considerably. Applications should therefore choose the appropriate confidence level for their own objectives.

Summing up, we conclude the following from the cross validation evaluation:

- The classifier is not over fitting to the training data;
- A very high precision is possible to obtain, but at the expense of recall;
- Recall never reaches very high levels, and lowers considerably if high precision is required.

5.3 Comparative evaluation of the Metadata Geoparser

To compare the results of the MG system against alternative solutions we have chosen to evaluate it against the Yahoo! Placemaker service. Placemaker was chosen because it is a commercial state-of-the-art system providing similar geoparsing functionality as the MG system. However, Placemaker does not support the geoparsing of metadata records so it can only be used as a black box geoparser.

This evaluation setup allowed us to evaluate our general approach. Since the Placemaker is prepared for geoparsing texts, if the MG system could outperform Placemaker than we would support our initial hypothesis that the context given by the metadata records can be exploited for performing place name recognition and resolution, and achieve better results.

For this comparison, the evaluation collection was processed by both systems and the results compared against those off the manual annotations. As before, the values of precision, recall, and F_1 were calculated. They are shown in Figure 5.

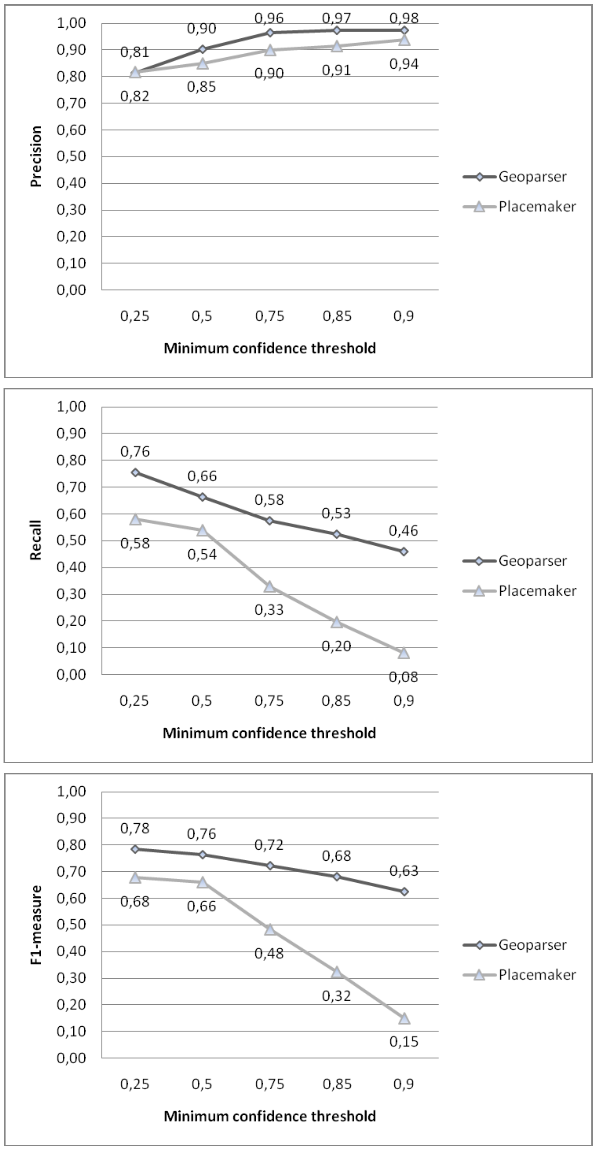

Figure 5 – Precision, recall and F_1-measure comparison between the MG system and the Yahoo! Placemaker

Analysis of the results shows that the MG system generally performed better than Placemaker. The MG system performed better in almost every measure and confidence level than Placemaker. The differences in the results between the two systems are statistically significant on all measures (p<0,001), except for the precision results on the lowest confidence level of 0,25, which were not statistically significant (p>0,5).

The superior results obtained by the MG system in this comparison, support our analysis from the cross-validation tests, that geoparsing involves a high level of uncertainty, and that the results for precision and recall are very positive.

The consistently higher results in precision and recall obtained by the MG system can be analyzed two aspects: the gazetteer behind each system, and the evidence used for recognition and resolution.

Placemaker uses Yahoo! GeoPlanet[7] as its gazetteer while the MG system uses Geonames. Both gazetteers seem similar in terms of comprehensiveness. They both contain data on more than six million places and our observation, during the manual annotation of the evaluation collection, was that none appeared to be more comprehensive than the other, although we did not measured it formally. The evidence the MG system used from data existing in Geonames (population, coordinates, administrative division hierarchy, and type geographic feature) is also available in GeoPlanet. So although we do not know the details of which data Placemaker uses from GeoPlanet, we don’t see any clear advantage of using Geonames over GeoPlanet.

Both systems are quite different in terms the evidence used for recognition and resolution. Although the implementation details of Placemaker are unknown to us, given its orientation towards the Web, it likely bases its processing on lexical evidence, and expects larger textual documents. The MG system uses very little lexical evidence, since it uses only word features and the token preceding the place name (see Section 4.2). In addition, the MG system also uses evidence from the metadata record, which Placemaker is unable to use.

Although we cannot exactly determine the factors that allowed the MG system to obtain better results, we believe that the main factors are in its independence of lexical evidence and the use of further evidence given by the structured data.

6. CONCLUSION AND FUTURE WORK

We presented and approach for place name recognition and resolution that best matches the characteristics of the unstructured text found in metadata records, and exploits relevant information from the structured data in these records.

The resulting system was designed to be independent of the languages and data format, as well as adaptable, by means of a machine learning technique.

The MG system is able to achieve very high precision of 0,99 at 0,55 recall, or reach a recall of 0,79 with 0,86 precision. Applications using the MG system can choose the appropriate balance between precision and recall for their purposes. They need to take in consideration that very high precision is possible to obtain, but with a higher decrease in recall.

The comparative evaluation, with an existing commercial service, showed that the precision and recall values obtained with the MG system were higher (p<0,001). Our analysis attributed the better results to two characteristics of the MG system: the independence of lexical evidence and the use of evidence given by the structured data.

We believe that the evaluation results support our general approach. Future work will concentrate on improving the results by researching the impact in the results of using other types of evidence, such as lexical evidence or other types of evidence from the structured data.

From the results of this work, it can also be hypothesized that our general approach is applicable to the recognition and resolution of other types of entities, such as persons, organizations, historical periods, etc.

[7] http://developer.yahoo.com/geo/geoplanet/

Future work should also address the machine learned classifier. State of the art techniques for named entity recognition use sequence labeling classifiers, namely Conditional Random Fields. Although we believe that sentences in metadata records are too short to get the benefits of these algorithms, future work will test and evaluate the use of this kind of classifiers.

7. ACKNOWLEDGEMENTS

This work was partially supported by the Europeana Connect project[8], which is co-funded by the European Commission programme eContentplus.

This work was partially supported by the Fundação para a Ciência e Tecnologia (FCT), through project grant PTDC/EIA-EIA/109840/2009 (SInteliGIS).

8. REFERENCES

[1] Seth, G., “Unstructured Data and the 80 Percent Rule: Investigating the 80%”, technical report Clarabridge Bridgepoints, 2008. <http://clarabridge.com/default.aspx?tabid=137&ModuleID=635&ArticleID=551>

[2] C. Shilakes, J. Tylman, "Enterprise Information Portals", Merrill Lynch report, 1998.

[3] S. Sarawagi, “Information Extraction”, Found. Trends databases, vol. 1, pp. 261-377, Now Publishers Inc., 2008. doi: 10.1561/1900000003

[4] Vatant, B., Wick, M., “Geonames ontology”. <http://www.geonames.org/ontology/>

[5] J. Leidner, “Toponym Resolution in Text”. PhD thesis, University of Edinburgh, 2007.

[6] Y. Kanada, “A method of geographical name extraction from Japanese text for thematic geographical search”, in proceedings of the 8th International Conference on Information and Knowledge Management, 1999.

[7] A. Olligschlaeger, A. Hauptmann, “Multimodal information systems and GIS: The Informedia digital video library”, in proceedings of the ESRI User Conference, 1999.

[8] C.J. Coates-Stephens, Sam. “The Analysis and Acquisition of Proper Names for the Understanding of Free Text”, Computers and the Humanities 26.441-456, San Francisco: Morgan Kaufmann Publishers, 1992.

[9] D. Nadeau, S. Sekine, “A survey of named entity recognition and classification”, Linguisticae Investigationes, volume. 30, number 1, pp. 3-26, John Benjamins Publishing Company, 2007.

[10] Y. Ravin, N. Wacholder, “Extracting Names from Natural-Language Text”, IBM Research Report , IBM Research, 1997.

[11] A. Mikheev, “A Knowledge-free Method for Capitalized Word Disambiguation”, in 37th annual meeting of the association for computational linguistics, Association for Computational Linguistics, pp. 159--166, 1999. ISBN:1-55860-609-3 doi:10.3115/1034678.1034710

[12] J. Silva, Z. Kozareva, J. Gabriel, and P. Lopes, “Cluster Analysis and Classification of Named Entities”, in proceedings Conference on Language Resources and Evaluation, 2004.

[13] D. Bikel, M. Daniel, S. Miller, R. Schwartz, R. Weischedel, "Nymble: a High-Performance Learning Name-finder", Proceedings of the Conference on Applied Natural Language Processing, Association for Computational Linguistics, 1997.

[14] A. McCallum, D. Freitag, and F. C. N. Pereira, “Maximum Entropy Markov Models for Information Extraction and Segmentation”, in Proceedings of the Seventeenth International Conference on Machine Learning, pp. 591-598, 2000.

[15] J.D. Lafferty, A. McCallum, and F.C.N. Pereira, “Conditional Random Fields: Probabilistic Models for Segmenting and Labelling Sequence Data”, Proceedings of the Eighteenth International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., pp. 282-289, 2001.

[16] B., Settles, “Biomedical Named Entity Recognition Using Conditional Random Fields and Rich Feature Sets”, in Proc. Conference on Computational Linguistics. Joint Workshop on Natural Language Processing in Biomedicine and its Applications, 2004.

[17] J. Pearl, "Bayesian Networks: A Model of Self-Activated Memory for Evidential Reasoning", in proceedings of the 7th Conference of the Cognitive Science Society, University of California, Irvine, pp. 329–334, 1985.

[18] M. Wick, T. Becker, “Enhancing RSS Feeds with Extracted Geospatial Information for Further Processing and Visualization”, in The Geospatial Web - How Geobrowsers, Social Software and the Web 2.0 are Shaping the Network Society, Springer, 2007.

[19] B. Christian, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, S. Hellmann, “DBpedia- A Crystallization Point for the Web of Data”, in Web Semantics: Science, Services and Agents on the World Wide Web, Volume 7, Issue 3, pp. 154-165, 2009. doi:10.1016/j.websem.2009.07.002

[20] E. Amitay, N. Har'El, R. Sivan, A. Soffer, “Web-a-where: geotagging web content”, in Proceedings of the 27th Annual international ACM SIGIR Conference on Research and Development in information Retrieval, 2004.

[21] Y. Kanada, “A method of geographical name extraction from Japanese text for thematic geographical search”, in proceedings of the 8th International Conference on Information and Knowledge Management, 1999.

[22] E. Rauch, M. Bukatin, K. Baker, “A confidence-based framework for disambiguating geographic terms”, in proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References, 2003.

[23] Borbinha, J., Pedrosa, G., Reis, D., Luzio, J., Martins, B., Gil, J., Freire, N. 2007. “DIGMAP - Discovering Our Past World with Digitised Maps“, in proceeding of the ECDL 2007 - Research and Advanced Technology for Digital Libraries, 11th European Conference, 2007.

[8] http://www.europeanaconnect.eu/

[24] A. Chandel, P.C. Nagesh, and S. Sarawagi, “Efficient Batch Top-k Search for Dictionary-based Entity Recognition”, Proceedings of the 22nd International Conference on Data Engineering, IEEE Computer Society, p. 28, 2006.

[25] J. Zhu, V. Uren, E. Motta, and J.Z.V. Uren, “ESpotter: Adaptive named entity recognition for web browsing”, 3rd Conference on Professional Knowledge Management, pp. 518-529, 2005.

[26] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan, “GATE: A framework and graphical development environment for robust NLP tools and applications”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002.

[27] Martins, B., Borbinha, J., Pedrosa, G., Gil, J., and Freire, N., “Geographically-aware information retrieval for collections of digitized historical maps”, in proceedings of the 4th ACM Workshop on Geographical information Retrieval, 2007.

[28] R. Bayer, E. McCreight, “Organization and Maintenance of Large Ordered Indices”, Mathematical and Information Sciences Report No. 20, Boeing Scientific Research Laboratories, 1970.

[29] S. Weibel, J. Kunze, C. Lagoze, M. Wolf, “Dublin Core Metadata for Resource Discovery”, Network Working Group Request for Comments: 2413, 1998.

[30] L. Breiman, "Random Forests". Machine Learning 45, Springer Netherlands, pp 5–32, 2001. doi:10.1023/A:1010933404324